Self-Monitoring Assessments for Educational Accountability Systems

Review draft: 26 June 2010

Self-Monitoring Assessments for Educational Accountability Systems

Daniel Koretz
Graduate School of Education
Harvard University
Cambridge, MA, USA

Anton Béguin
Cito
Arnhem, the Netherlands

To appear in *Measurement: Interdisciplinary Research and Perspectives, 8*(2-3), 2010.

Review draft: 26 June 2010

Abstract

Test-based accountability is now the cornerstone of U.S. education policy, and it is becoming more important in many other nations as well. Educators sometimes respond to test-based accountability in ways that produce score inflation. In the past, score inflation has usually been evaluated by comparing trends in scores on a high-stakes test to trends on a lower-stakes audit test. However, separate audit tests are often unavailable, and their use has several important drawbacks, such as potential bias from motivational differences. As an alternative, we propose *self-monitoring assessments* (SMAs) that incorporate audit components into operational high-stakes assessments. This paper provides a framework for designing SMAs. It describes five specific SMA designs that could be incorporated into the non-equivalent groups anchor test linking approaches used by most large-scale assessments and discusses analytical issues that would arise in their use.

Review draft: 26 June 2010

In recent decades, test-based accountability (TBA) has become the cornerstone of U.S. education policy. Pressure on educators to raise scores has increased from one wave of reforms to the next. TBA, well-established for some time in the U.S. and England, is now appearing in many other nations as well.

The net effects of these TBA policies, and particularly, variations in the net effects across types of schools, students, tests, and accountability systems, remain uncertain. However, research has made clear that in their attempts to raise scores, educators often resort to undesirable strategies. These include focusing too narrowly on tested content and providing test preparation that capitalizes on substantively unimportant aspects of the test, such as format or unimportant features of scoring rubrics (e.g., Stecher 2002, Stecher & Mitchell, 1995). These responses can undermine the test's representativeness of the domain and thereby produce score inflation, i.e., increases in scores that are larger than improvements in mastery of the domain would warrant. Research has shown that this inflation can be very large (Klein, Hamilton, McCaffrey, & Stecher, 2000; Koretz & Barron, 1998; Koretz, Linn, Dunbar, and Shepard, 1991).

These distortions should not be surprising. They are a manifestation of Campbell's Law:

> The more any quantitative social indicator is used for social decision making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor (Campbell, 1979, p. 87).

The distortions described by Campbell have been documented in a wide variety of fields other than education (e.g., Rothstein, 2008).

As Feuer (this volume) and others have argued, Campbell's Law is not in itself reason to avoid performance-based accountability systems. Even with distortions, the net effect of such a system can be strongly positive. However, the presence of severe distortions, such as have been found in research on test-based accountability programs, does warrant response. Feuer (this volume) notes two responses: better evaluation of the effects of the program, and efforts to design the programs to minimize unwanted effects and maximize positive effects.

There are two distinct but overlapping strategies for better design. The first is to design testing programs to make them less vulnerable to undesirable behavioral responses and score inflation. The second is to design the accountability programs into which tests are embedded to lessen such responses, for example, by giving substantial weight to factors other than increases in standardized test scores e.g. judgments from an inspectorate or non test based indicators, such as percentage drop-outs. In this paper, we address only the former. We suggest an approach to test design that should both decrease the incentives to prepare students inappropriately and facilitate better evaluation of the effects of accountability, both positive and negative.

Evaluating Score Inflation

In most of the research to date, score inflation has been evaluated by comparing trends on a high-stakes test to trends on an audit test—a low- or lower-stakes test intended to measure a reasonably similar domain of achievement. In the U.S., the National Assessment of Educational Progress (NAEP) has been used most often as an audit because it is a broad measure, reflects a degree of consensus about the goals of education, and is rarely the focus of explicit test preparation. However, some districts and

states administer a second, lower-stakes test, and these have been used as audit tests in a few studies (Jacob, 2005; Schemo & Fessenden, 2003).

This approach has several important limitations. Suitable audit tests are often unavailable or are infrequent. For example, NAEP is administered in only three grades, and not every year. When data from a second test are available, its suitability as an audit may be limited. For example, the substantive appropriateness of the second test may be arguable. Even when their substantive appropriateness is clear, audit tests are generally not on the same scale as the high-stakes tests to which they are compared. In addition, if students know that the audit test has no consequences, the comparison may be confounded with motivational effects. While differences in motivation are less problematic for comparisons of trends than for cross-sectional comparisons, they are nonetheless potentially a problem for the former as well. Student-level exclusions may be different for the high-stakes and audit tests—and may change differentially over time—biasing comparisons between them. When audit tests are administered only to samples of schools, there is a risk of accidental or intentional differences in samples over time. Creating and administering a new audit test rather than relying on extant measures would address only some of these limitations.

We propose an alternative to separate audit tests: *self-monitoring assessments* (SMAs). SMAs would incorporate one or more audit components into the operational forms of high-stakes tests. In some but not all cases, this approach would allow the audit and high-stakes measures to be placed on the same scale, and it would address the other limitations of separate audit tests noted above as well.

This paper describes methods for designing and using SMAs. The following section briefly sketches the framework for evaluating validity of inferences under high-stakes conditions detailed by Koretz and colleagues (Koretz, McCaffrey, and Hamilton, 2001; Koretz and Hamilton, 2006), which undergirds our design for SMAs. The paper then sketches several different designs for SMAs. A subsequent section explores some technical issues that these designs raise for analysis. A final section discusses implications and unresolved issues.

A Psychometric Framework for SMAs

In the United States, the evolution of test-based accountability has contributed to a number of innovations in test design. These include the development of a variety of performance assessments and other approaches for cognitively richer assessments; the development of 'standards-based' tests aligned with states' content and performance standards; innovations in methods for setting performance standards; and advances in growth modeling. However valuable, none of these developments directly addresses what we consider to be the core problem underlying score inflation: predictable recurrences of substantive and nonsubstantive sampling in the design and construction of tests.

To evaluate the potential for score inflation, it is necessary to specify what aspects of sampling can offer the potential for inappropriate test preparation. Koretz et al. (2001, 2006) suggest that for evaluating the validity of inferences under high-stakes conditions (VIHS), we view a test as a collection of *performance elements*, a deliberately vague term that denotes all aspects of performance that contribute either to performance on a test or to the inferences about achievement that are based on it. *Substantive* elements are those that are relevant to the inference based on scores. *Non-substantive* elements are not the

focus of inference. Koretz and Hamilton (2006, p. 545) give the example of "facility with a particular format that is of no particular importance for the intended inference." It is important to note, however that non-substantive elements include far more than item format, which is typically used to refer only to multiple-choice, short constructed-response, and so on. For example, some problems in elementary algebra may be represented verbally, algebraically, graphically, or pictorially. This choice, which may be of no substantive importance, may be used repeatedly. A test writer may consistently present graphs of linear equations in one variable with positive slopes and positive intercepts, even if the signs of the slope and intercept are of no substantive importance for the inference. Therefore, it is more accurate to refer non-substantive elements as aspects of *item style* rather than format.

Response demands may also be both substantive and non-substantive. Scoring rubrics often provide opportunities for coaching that Stecher and Mitchell (1995) called "teaching to the rubric"—as described by one teacher, "What's in the rubrics gets done, and what isn't doesn't." Stecher and Mitchell noted that this "May cause teachers to neglect important…skills not addressed by the rubrics and neglect tasks not well aligned to [them]" (1995, p. ix). The choice among rubrics, however, is often at least in part non-substantive.

All of these substantive and non-substantive design decisions can affect scores. Koretz et al. (2001) define an element's *effective test weight* as the sensitivity of the score to change in performance on that element. They assume no specific model for compositing performance on elements into scores. If the test score ζ is any function of the performance elements π_i,

(1) $$\zeta = f(\boldsymbol{\pi}),$$

then the effective test weight of element *i* is simply the partial derivative:

(2) $$\lambda_i = \frac{\partial \zeta}{\partial \pi_i}.$$

The inference based on scores defines the latent construct to be measured and is called the target of inference. Each performance element also has an *inference weight*, which indicates its importance to the target. Inference weights are generally vaguely and incompletely specified, apart from initial test blueprints, and they cannot be readily formalized mathematically. Nonetheless, they correspond logically to test weights, and that correspondence is the key to VIHS.

One can represent a test and the target of inference as two vectors of weighted performance elements. The two vectors together comprise three sets of elements. The first, tested elements, set comprises elements that are relevant to the inference and that appear in the test, albeit not necessarily with equal weights in both. The substantive portion of this first set is typically the focus of discussions of alignment. For example, in a test examined by one of the authors, coordinate geometry was given far more weight than was warranted by the target suggested by state content standards because it was often used incidentally in items assessing algebra. The second set comprises *implicit elements*: performance elements that are relevant to the inference but that are omitted from the test. When inferences are made about broad domains, as is commonly the case in TBA systems, the set of implicit elements is necessarily very large. It is important to note that tabulations of the proportions of a state's standards that are assessed do not address the size of the implicit set. Most standards can be represented in many different

ways, both substantively and nonsubstantively. Therefore, most relevant performance elements may remain untested even if the testing of content standards is exhaustive.

Finally, a third set of elements are those that affect performance on the test but are irrelevant to the inference and therefore have inference weights of zero. In constructing a test, it is necessary to make a great many decisions about item style and response demands, some of which are substantively unimportant. These choices are represented by performance elements in this third set.

Koretz et al. (2001) suggest that validity of inferences under any circumstances can be expressed as the degree to which performance on tested elements, as weighted by the test, warrant inferences about mastery of the target, including untested elements, as weighted by inference weights. Thus, validity can be seen as a matter of generalizability, consistent with Kane's sampling model of validity (Kane, 1982, 2006).

Under low-stakes conditions, test developers are concerned primarily with the size and *initial* representativeness of the tested sample from the larger target. If the sampling is appropriate and stakes are low, the primary consequence of incomplete sampling is measurement error. The effects of multidimensionality, manifested, for example, in differences in results across alternative tests, are typically modest because of the generally high cross-sectional correlations among subsets of the target. This correlation between subsets of the target allows generalization among them and makes the test score a good representation of the score on the total domain.

VIHS, however, poses additional challenges. High stakes can induce educators and students to focus unduly on the tested elements, both substantive and non-substantive. This can generate improvements in performance on tested elements that are

not reflected in performance on elements that are either untested or given very low test weights. Regardless of whether the emphasized elements are substantive or not, these behaviors and the resulting performance undermine the test's representation of the domain and therefore inflate scores. This process will often entail an exacerbation of multidimensionality, as performance on the elements emphasized on the test changes independently of latent performance on unmeasured elements. Gains in scores will no longer accurately predict changes on the elements of the target that are not measured in the test. As a consequence, generalizations from the tested set to the target of inference will be biased.

Koretz et al. (2001, 2006) suggest classifying test preparation activities—including all methods used to prepare students, whether desirable nor not—in terms of this model of test construction. Leaving aside cheating, they suggest two forms of test preparation that can, by different mechanisms, inflate scores. *Reallocation* refers to shifts in instructional resources, such as instructional time, to better match the specific sampling of substantive elements used to construct the test—for example, eliminating topics with zero test weights in order to spend more time on topics with large test weights. Reallocation occurs both across subject areas and within them (e.g., Stecher, 2002). Reallocation within a tested subject causes score inflation when the elements receiving less emphasis have substantial inference weights. Reallocation does not bias estimate of performance on individual elements, but it does bias the test score by undermining the ability of the tested set of elements to represent the larger set that has substantial inference weights.

Koretz et al. use the term *coaching* to refer to an instructional focus on details of the test that are not important to the inference. Non-substantive coaching focuses on non-substantive elements. An example would be advising students to use process of elimination to 'solve' mathematics problems presented in the multiple-choice format. Another example, which appears in a test-preparation book written for the 10th grade Massachusetts MCAS test, takes advantage of the fact that items testing knowledge of the Pythagorean theorem will generally require integer solutions and advises students simply to use the "popular Pythagorean ratios" of 3:4:5 and 5:12:13 and their multiples (Rubentstein, 2002, p. 56).

Coaching may focus on substantive elements as well, but on substantive details fine-grained enough that the distinctions among them are unimportant for the inference. For example, the test-preparation cited above notes predictable patterns in the plane geometry items included in the test. Several of these are labeled "special triangle rules," of which the first is this: "One triangle rule that is often tested on the MCAS exam is the *third side* rule. The rule is: The sum of every two sides of a triangle must be greater than the third side" (Rubinstein, 2002, p. 52). Unless the selection of that specific bit of content to represent the performance element—in preference to many others on related topics in plane geometry—is important enough for the target to have its own nonzero inference weight, focusing attention on this particular content would constitute substantive coaching. Although the dividing line between reallocation and coaching is sometimes vague, they are in theory different: coaching causes inflation not only by biasing the aggregate score, but also by biasing estimates of performance on individual elements.

Non-substantive coaching, when it works, induces what construct-irrelevant performance variation. As Messick put it:

> Construct-irrelevant easiness occurs when extraneous clues in item or test formats permit some individuals to respond correctly in ways irrelevant to the construct being assessed (Messick, 19819, p. 34).

In the case of VIHS, however, we need to worry not only about "extraneous clues," but also about direct efforts by teachers and others to instruct students in ways of capitalizing on these incidental recurrences. In contrast, the bias caused by reallocation does not fit the traditional notion of construct-irrelevant variance. In the case of reallocation, certain *relevant* performance elements are given additional emphasis, but improved performance on those elements may be free of bias. It is the aggregation of these elements into a score that represents the target of inference that introduces bias.

The substantive and non-substantive recurrences in sampling that provide opportunities for inflation-producing test preparation are common, and they can be quite extreme. Items in one a given form often resemble very closely those in previous forms in both content and item style. (For a few example of near-clone items, see Eduwonkette, 2008). These recurrences may arise for many reasons. In some cases, they may be unintentional, a reflection of constraints of time, money, or imagination. However, recurrences may also be intentional. For example Morley, Bridgman, and Lawless (2004) describe efforts by the Educational Testing Service to create item models to facilitate the development of highly similar items:

> An item model can be thought of as a means of generating close variants with the intention that the isomorphs will be psychometrically and otherwise exchangeable and equivalent (p. 1).

Dunbar (2008) notes that the evolution detailed grade-level expectations has exacerbated this problem, in some cases resulting in item shells that can be directly coached. While recurrences have the advantage that they lessen unwanted differences in performance that can introduce noise into linking across forms, they have a high price in a high-stakes context. Test-preparation materials often focus explicitly on these recurrences in order to facilitate inappropriate test preparation, and many educators identify them independently for the same purpose.

Design Principles for SMAs

Score inflation can have a variety of causes, including manipulation of the tested population (e.g., Figlio & Getzler, 2006; Jacob, 2005) and cheating as well as reallocation and coaching. SMAs are intended to address the inflation caused by reallocation and coaching, but they might also address some forms of cheating.

The fundamental principle underlying the design of SMAs is to incorporate *audit items* that would be less susceptible to inflation and hence could be compared to more vulnerable items in the operational assessment. Audit items could be of several sorts. One class, which could be called *content audit items*, would assess content important for the inference that is not tested by the operational item sample. For example, Holcombe, Jennings, & Koretz (2010) show persistent gaps in the coverage of standards by some state tests. In such cases, substantive audit items would assess content not covered by operational items. The second class of audit items, *style audit items*, would address

possible inflation from predictable non-substantive recurrences. For example, using GRE items, Morley, Bridgeman, & Lawless (2004) showed that what they called alternatively "close variants" or "isomorphs"—mathematics items that shared item style as well as content with other items—are in general easier for examinees than "matched items" that share mathematical content but not item style. In this case, what they dubbed "matched items" could serve as style audit items. Consistency in performance between the operational and audit items would be convergent evidence of validity. Discrepancies between operational and audit items—which might be cross-sectional differences in performance or differential divergence over time—would signal possible score inflation.[1]

The essential first step in either identifying or creating a suitable audit measure is to clarify the target of inference. This may seem trite, but in practice, there has been considerable disagreement about this, stemming in part from an excessive focus on the characteristics of the tests themselves rather than the target. For example, shortly after No Child Left Behind was enacted, the National Assessment Governing Board convened a panel to consider whether and how NAEP should be used as an audit for the state tests used for accountability under the new law. The panel concluded that "any amount of growth on the National Assessment should be sufficient to 'confirm' growth on state tests" because there are a variety of factors that could limit consistency in trends between

[1] The findings of Morley et al (2004) suggest a third possible form for audit items. They found that "appearance variants," items that shared item style but not content with comparison items, were in general more difficult than either matched items or isomorphs. The implications of this for present purposes are unclear. If the content of the appearance variants was sufficiently different from that of comparison items, the appearance variants would be content audit items. However, it is also possible that the weak performance on these items represents a focus on item style that distracts students from content. If so, this would provide an opportunity for a third type of audit, which would be harder rather than easier than comparison items in the presence of score inflation. This possibility is not addressed further in this paper.

the two tests, including format and difficulty (Ad Hoc Committee on Confirming Test Results, 2001, p. 9). Therefore, they argued, discrepancies should be seen as a threat to validity only "when there is consistent, compelling contrary evidence from the National Assessment that *cannot be explained simply by the differences between the two tests* or other relevant factors" (p. 9, emphasis added). This argument mistakes the design of the accountability test for the target. Format, for example, is in most cases a non-substantive performance element, and therefore, differences in performance attributable to format are generally a threat to validity (see Koretz, 2007).

Once the target has been clarified, the second step is to identify actual or potential recurrences that might threaten VIHS. This requires examination of both operational test forms and scoring rubrics. Examination of test-preparation materials can also be helpful, not only in providing information about recurrences, but also by pointing out recurrences that teachers are particularly likely to know about and that therefore may be especially valuable for designing the SMA.

Recurrences of performance elements, even very minor ones, will often be apparent when multiple forms of a test are compared. In contrast, repeated omissions or underweighting—that is, assigning less emphasis on the test than inference weights warrant—is likely to be more difficult. Omissions or underweighting of large portions of content, for example, specific content standards, can be readily ascertained. In contrast, omissions and underrepresentation of more fine-grained elements can only be identified by keeping the large range of relevant elements in mind. However, both underweighting and omissions are essential for reallocation-based score inflation because teachers need to move resources away from performance elements with low test weights, and surveys

indicate that many teachers do reduce emphasis on material that they have identified as omitted from or deemphasized by the test (e.g., Koretz, Baron, Mitchell, and Stecher, 1996).

SMA Designs and Analytical Ramifications

In this section, we describe several SMA designs and note some of their advantages and disadvantages.

The SMA designs described here differ on two related dimensions. The first dimension is timing: whether the audit items are included with the initial administration of the operational assessment or are introduced later. The second dimension is the analytical approaches they allow.

Analyzing two or more waves of assessment data poses an identification problem: there is more than one set of parameters that can generate the data. In current practice, his indeterminacy is typically resolved by accepting one scaling model and fixing the scale of the first administration arbitrarily. With these constraints, the identification problem is resolved. Linking—in current programs, most often a non-equivalent groups anchor test (NAT) design—allows one to place subsequent waves onto the same arbitrary scale. Such approaches require two assumptions. First, observed performance on the linking items is interpreted as a representative measure of the proficiency, although one subject to sampling error (e.g., Fitzpatrick, 2008). Second, the relationship between the proficiency and observed performance on the linking items is assumed to remain constant, up to the linear transformation of scale that linking removes.

Neither of these assumptions is warranted under high-stakes conditions.

Inappropriate test preparation may make linking items appear easier than true changes in proficiency warrant (Koretz, 2007). This undermines both of the required assumptions.

Some SMA designs permit evaluation of the assumptions of the linking procedures and can allow for placement of both conventional and audit items on a single scale. However, in other cases, the identification problem is not soluble, and it is not possible to place all items in more than one wave of data on a single scale. For this reason, some SMA require analytical methods that do not depend on a single scale, but instead evaluate differences-in-differences between parts of the design.

Graphical representations are provided for the SMA designs, in which items are on the horizontal axis and persons on the vertical axis. Blocks of items indicate which items are administered to which persons. In the figures, operational, linking, and audit items are portrayed as distinct blocks for simplicity of presentation. The discussion here focuses on comparisons of these blocks. A more powerful approach would pair audit items to the specific operational items that share the characteristics that generated them. Paired-item approaches are not discussed further, but the analytical issues described below generalize to them.

For clarity, these diagrams are simplified in several respects. In practice, the items of a given type, such as audit items, would most likely not be administered in intact blocks. The size of the blocks in the diagrams is not necessarily representative of the number of items. Moreover, operational versions of the designs might be more complex. For example, to limit the number of items administered to a single person, it may be necessary to use incomplete versions of the designs. In that case, some students would take only operational and linking items, some would take operational and audit items, and

other would take operational items and a proportion of both the linking and audit items. Although this would add to the complexity of the analysis it would not change the fundamental characteristics of the designs described below or the central analytical issues we address below.

Initial Incorporation into the Operational Assessment, Static Audit (IIS Design)

In this approach, audit items are incorporated into the test from the first implementation of the assessment, and operational and audit items are scaled together from the outset. Figure 1 shows an IIS design that is an adaption of a standard NEAT linking. In this example, 2009 is the first operational administration of the assessment. *O* signifies operational items not used for linking, and the numeral is the year of administration. *L* denotes conventional linking items administered in both years. *A* signifies audit items.

Figure 1 about here

In this design, the linking and audit items differ in purpose and therefore in design and content. The linking items provide the conventional estimate of aggregate growth on the operational tests. If performance on the linking items L1 differs in the two years, this will be modeled as a difference in proficiency between the population in 2009 and 2010. If no difference occurs on the linking items and performance on the operational items O2 is better than on items O1, this will be modeled as a difference in item difficulty. This will affect the relative scores of individuals and subgroups, e.g., schools, but it will not contribute to the difference in performance of the entire group in 2010 relative to the

group in 2009. The audit items aim at identifying inappropriate reallocation and coaching, which will create differences of response behavior between the audit items and operational and linking items.

In the IIS design, the common calibration of items in the first administration solves the identification problem. This approach simplifies scaling, eliminates the need for linking other than that done in the operational assessment, and maximizes and simplifies options for analysis because the audit and operational items are on the same scale. The audit items can be used as an anchor for subsequent DIF analysis, and differences in performance are directly interpretable.

However, the disadvantages of this approach are also substantial. First, the recurrences that provide opportunities for inappropriate reallocation and coaching may not be apparent at the start of a testing program. It is likely that much of this recurrence are initially unplanned, and much of it will not be apparent from examination of the initial forms alone. Moreover, it will not be apparent initially which of the opportunities for reallocation and coaching educators will take advantage of. This makes the design of audit items difficult and may make the IIS design vulnerable to Type II errors, that is, failing to show inflation when it has occurred. The smaller and less comprehensive the audit section, the more difficult it would be to lessen this risk.

A second potential disadvantage of the IIS design is that it removes the element of surprise. Inflation is facilitated when educators or students—or those coaching them—identify likely recurrences in the operational assessment. In the IIS design, the audit items are present from the outset and therefore may be identified by those looking for potential recurrences. This risk is lessened to the degree that the audit items are designed not to

share obvious characteristics of other items because there is little potential gain from identifying possible low-frequency recurrences. The risk of coaching on the audit items might also be limited by administering the audit items to a sample of respondents instead of the total population in the first year. This would preserve the element of surprise for most of the tested population in the second year.

Given these disadvantages, the most likely use of an IIS design would be in response to planned incompleteness in the test, e.g., a design decision that portions of certain standards will be de-emphasized or that specific item styles will be employed for the bulk of the test.

Initial Incorporation into the Operational Assessment, Dynamic Audit (IID Design)

An alternative model of initial incorporation of an audit block requires calibrating a number of audit blocks together then alternating their use over time. Figure 2 shows an example of this design using NEAT linking and chain equating. In this example, all of the audit blocks are calibrated together with the first operational assessment, but other approaches might be followed to link the audit items to the scale of the operational assessment. The primary advantage of the IID design over the simpler IIS is that it lessens the risk of inflated performance on the audit items. To achieve this goal, however, the audit blocks must differ in terms of incidental characteristics of items that might facilitate inappropriate test preparation. Because it can make use of a broader and more diverse item set across all the audit blocks, the IID may be somewhat less vulnerable than the IIS design to missing some of the focuses of test preparation and therefore underestimating score inflation.

Figure 2 about here

Post-hoc Incorporation, NEAT Linking Only (PIN Design)

In this design, the audit component is added after two or more years of operational administration. Figure 3 shows a PIN design with NEAT linking and chain equating. The audit items might be added at any time, but in this example, they are added in the third year. In a modification of this design, additional audit blocks could be added in subsequent years.

Figure 3 about here

The primary advantages of this design mirror the primary disadvantages of the IIS and IID designs: PIN allows the test designers to look for replications in the operational test forms and to acquire information on publicized test-preparation strategies from test-preparation materials and primary data collection in schools (either surveys or direct observation). This should greatly increase the power of the audit.

However, the analytical costs of using the PIN design rather than the IIS or IID design are large. Without additional data, the identification problem returns for the first three years: one cannot be confident that conventional linking methods successfully place the audit items and the operational items of the first years on a common scale. One could calibrate the items together, but the results of this common calibration might not have the usual and desired interpretation. For example, if a NEAT linking design is used in the operational assessment, inappropriate test preparation may generalize to the linking

items, as a result of which their difficulty will decline more than change in latent proficiency warrants. Linking fixes the difficulty estimates for these items when in fact the parameters have shifted, and the result is that the proficiency scale will be shifted upward by an unknown amount (Koretz & Barron, 1998; Koretz, 2007). If the difficulty of new audit items is estimated by calibrating them with the linking items, these estimates will be biased upward as well. That is, the audit items will be more difficult relative to the linking items than they should be, and that places the audit items too high on the original scale. The only solution to this problem is to use additional data, such as an external anchor, to resolve the indeterminacy in the difficulty estimates of the audit items.

As a result, the PIN design may underestimate inflation in the population as a whole, including the large inflation that often occurs during the first several operational administrations of a new test. However, PIN does allow one to examine *variations* in inflation cross categories of schools or classrooms. For example, inflation might be expected to vary as a function of the demographic characteristics or achievement profiles of schools, or as a function of rate of change in scores on the high-stakes test. A PIN SMA would be well-suited to evaluating such patterns because doing so does not require having the operational and audit components of the SMA on the same original scale.

Combined Initial- and Post-hoc NEAT Design (IPIN Design)

If one had enough space in an operational assessment, one could combine the PIN design with the IIS or IID design. An initial set of audit items would be incorporated into the operational assessment in the first administration, and an additional set would be added after some time, perhaps two years, to capture the effects of specific replications

and test-preparation strategies. Figure 4 shows an IPIN design, adding an audit component, starting in the first year, to the design shown in Figure 4.

Assuming that the original audit items (A1) are not vulnerable to inflation, one can put the later sets of audit items onto the initial scale using the link through these items. To check this assumption additional data, such as an external link, can be used to put the later sets of audit items on the initial scale. Therefore, the IPIN Design provides more analytical options than does a simple PIN design.

_____

Figure 4 about here

_____

Post-hoc Incorporation, External Common-Persons Linking (PIEC design)

All of the preceding designs assume a common-items linking strategy. Of these, only the IIS, IIID and IPIN designs solve the identification problem.

Under high stakes conditions, the alternative approach to identification is to rely on common-persons linking, which circumvents the problem of the shift in item difficulty described above. For purposes of an SMA, students from a jurisdiction not subject to the specific high-stakes test would be administered both operational and audit items to place them on the same scale. Figure 5 shows one possible PIEC design. The operational assessment follows a typical NEAT design. In the second year, an audit block is administered in another jurisdiction, along with the first linking block, and these items are calibrated together to place them on the same scale. A number of conventional techniques could be used to align this scale with the original operational scale, such as

constraining the parameters of the *L1* items to equal those obtained in the first operational assessment.

Figure 5 about here

The PIEC design has the advantage that the external anchoring can be done at any time, and therefore, it can avoid the primary weakness of the PIN design: it allows us to put the audit component and the operational component on a single scale even if the audit component is added after the first year of testing. However, this approach has also some serious drawbacks.

First, this approach is more expensive and substantially more burdensome than the previous ones, and it this requires cooperation from administrators and teachers in other jurisdictions who have little to gain by participating in linking studies. The amount of testing already required by high-stakes testing systems may make this participation hard to obtain.

Second, any approach that relies on external anchor testing requires population invariance in the relationships between operational and audit item difficulties across the linking target jurisdictions. In practice, this assumption can only partly be tested by evaluating item drift between administrations in the different populations, and the assumption may not be entirely reasonable, for example, because of differential match with curricula in the two jurisdictions.

Third, there is a risk of motivational effects, particularly in the older grades, a risk that is exacerbated by the amount of testing required in some high-stakes systems. A

motivational effect that is uniform across characteristics of items would not be a problem, but anecdotal evidence suggests that in some cases, the negative motivational effect is on average larger for extended constructed response items than for many multiple choice items. Indeed, in one high-school level field-test pre-equating effort with which we are familiar, most students simply did not respond to difficult constructed response items. Because the differences between operational and audit items may create differences in difficulty—or differences in perceived difficulty, which is almost as serious—the results obtained in the external linking sample may be misleading. For example, the audit items for multiple-choice problems that are vulnerable to process of elimination would likely be constructed response and therefore might seem harder and generate lower levels of effort in the linking study. This would yield an upwardly biased estimate of the difficulty of the audit items and a downwardly biased estimate of inflation when the audit items are used in the target jurisdiction. One might lessen this risk by incorporating the linking study in a moderate- or high-stakes assessment, but this may be an unappealing option for host schools, given the stress it might create among teachers and examinees.

Analytical framework

The goal of the analysis of an SMA is to evaluate whether the performance on audit items is similar to performance on the operational items of the assessment. At the highest level of aggregation, the population level, the comparison between audit items and operational items will entail only a comparison of behavior on linking and audit items. The (non-linking) operational items of different administrations are not directly comparable to each other, and they are irrelevant to population-level estimates of change. In contrast, direct comparisons between audit and operational (non-linking) do provide

useful information about the relative performance of smaller aggregates, such as schools, as described below.

We distinguish between two different analytical approaches, the *scale-based approach* and the *evaluation of differences*. The scale-based approach relies on calibrating the items to a single scale and identifying differential item functioning (DIF) or item drift between the linking items that could have been coached and the audit items that are not coached. However, in some of the SMA designs, coaching could have introduced item drift in the linking items, and therefore the identification problem is not solved. The evaluation of differences approach entails looking at higher-order differences—for example, comparing the performance difference between audit and operational items across categories of schools. This approach can be applied with any SMA design, including those that do not solve the identification problem.

The scale based approach at the population level

The first step in the analysis is to establish a benchmark that describes the relationships between the characteristics of the regular items and the audit items if there is no coaching on the regular items. We will illustrate this with the IIS design, which is the simplest case, but similar approaches could be followed with the IID, PIN and IPIN designs.

In the IIS design as illustrated in Figure 1, the benchmark is obtained by calibrating the 2009 audit items (A1), operational items (O1), and linking items (L1) together (although as noted earlier, at the population level, only (L1) and (A1) matter for this purpose. In this case, the calibration is done in the first year, before there has been an opportunity for inflation.

This approach can be described in terms of either classical test theory or item response theory (IRT) models (Lord, 1980; Fischer & Molenaar, 1995; VanderLinden & Hambleton, 1997), but in this section, we describe them in terms of a three-parameter logistic IRT model for binary items. Let x_{ij} be a random variable indicating the response of an individual *j* on an item *i,* with values $X_{ij} = 1$ for a correct answer and $X_{ij} = 0$ for an incorrect answer. In this model, the probability of correct response is given by:

$$P(X_{ij} = 1) = \gamma_i + (1 - \gamma_i)\frac{\exp(\alpha_i(\theta_j - \beta_i))}{1 + \exp(\alpha_i(\theta_j - \beta_i))} \quad (3)$$

where the item parameters can be interpreted as the difficulty β_i, the discrimination α_i and the probability of a correct guess on the item γ_i. The parameter θ_j can be interpreted as the proficiency of the person. A unidimensional IRT model of this sort rests on the assumption that systematic differences in behavior on two or more items can be described fully by differences in the item parameters and that a single parameter describes the proficiency of the person. Consequently if the proficiency of a person is higher, this has an effect on the response behavior and the probability of a correct response on all items. Based on the linking of the 2009 data (O1, L1 and A1), the parameters of items in all three blocks are placed on a single scale.

Inappropriate test preparation may improve performance on operational and linking items more than improvement on the relevant underlying construct warrants. The design of SMAs is based the assumption that to the extent that the audit items avoid the performance elements responsible for this bias, performance on the audit items will not be biased. Having both biased and unbiased items in the assessment would violate of the assumptions of the IRT model because it would introduce a systematic source of

variations in performance that is not captured by the single person parameter, θ. This would appear as differential item functioning (DIF). Specifically, at the population level, we expect DIF in the form of differential changes in the item parameters of the linking items, compared to those of the audit items, in the second or subsequent administrations of the SMA.

The first step in analysis of potential inflation is to evaluate whether this differential growth occurred. This can be assessed, for example, by comparing the fit of the unidimensional IRT model in which every item has the same parameters in 2009 and 2010, with a more general model that allows for different item parameters for the L1 items in 2009 and 2010. The two models can be evaluated using global fit statistics like the BIC and AIC or with a number of other test statistics that can be designed (see, e.g., Glas and Verhelst, 1995). Both general inflation and drift of specific item parameters can be identified. A confirmatory approach is to evaluate whether items already suspected of being coachable (e.g., Holcombe et al., 2010) show atypically large drift. An exploratory approach, which one might confirm with cross-validation, would identify items showing the greatest drift and then evaluate their substantive and non-substantive characteristics (Koretz & McCaffrey, 2005). As an alternative to concurrent calibration of the 2009 and 2010 data one can use separate calibration. Firstly, item parameters are established on the 2009 data. Then item parameters in the ISS design are estimated using the 2010 data (part O2, L1 and A1 of the design). In this way, one can also evaluate whether differential growth occurred on specific linking items compared with the audit items.

The analysis of the PIEC design (Figure 5) follows the same logic, but with one difference. In the PIEC design, the (L1) and (A1) blocks are not necessarily administered

in the first year of the operational assessment, and the design depends on a common-persons linking in another jurisdiction to place these blocks on a common scale.

The scale-based approach at a lower level of aggregation

Similar procedures can be applied to groups at a lower level of aggregation. At lower levels of aggregation, however, there is additional information available. While only the linking items contribute to estimates of population growth, all operational items contribute to the cross-sectional dispersion of scores, including variability among aggregates such as schools. Therefore, one can make use of relative differences in performance between audit and non-linking operational items.

The evaluation of differences approach

Some of the SMA designs cannot provide a common scale in all parts of the design. We will use the PIN design (Figure 4) to illustrate this case because we expect that it will be one of the most common applications of an SMA approach. In this case, we assume an operational assessment that uses IRT NEAT linking across years, to which audit items are added only after the operational assessment has been administered at least one previous time. In this discussion, we assume that the audit items are first added in the third administration, so that educators and test preparation firms have an opportunity to examine recurrences in the operational test before choosing test preparation approaches. However, the length of time before the addition of audit items is immaterial to this discussion, as long as they are not incorporated into the initial administration.

The lack of a common scale in the PIN design arises from a violation of a necessary assumption of NEAT linking. This assumption is that the relationship between θ and the difficulty of items in the linking item is constant across administrations.

Differences in the values of $\hat{\beta}$ obtained in two adjacent years reflect only a meaningless linear transformation of scale that is removed by linking. However, if inappropriate test preparation makes the linking items easier than changes in θ warrant—either directly, or by generalization from preparation for similar items not used in linking—this assumption is false. The linking transformation intended to remove the meaningless linear transformation between the two scales will instead build score inflation into the later scale. Therefore, in the PIN design, the appropriate scale for auditing cannot be identified.

To make this concrete, consider the PIN design in Figure 4. In 2011, the difficulty of the O3 and L3 blocks relative to the original scale is unknown by design, the difficulty of block L2 is unknown because of the risk of inflation, and the difficulty of the new audit block A is unknown because it was never administered in conjunction with the other blocks under circumstances in which their difficulty relative to the original scale is known. The initial discrepancy in observed performance between the audit and other items (called the *performance discrepancy* here) expected in the absence of inappropriate test preparation and score inflation is unknown.

Thus, while score inflation after the introduction of the audit items can be observed directly—because DIF arising after that time will be apparent—overall inflation up to the introduction of the audit items cannot be observed directly. One analytical solution to this indeterminacy is to use higher-order differences. That is, one would look for systematic relationships between the performance discrepancy and other variables on the assumption that it should be greater in schools that do a larger amount of successful

but inappropriate test preparation. For such comparisons to be informative, it is not necessary for the audit items to be on the same scale as the operational or linking items.

Koretz & Barron (1998) followed this logic but using only linking and operational items in a study of Kentucky's KIRIS testing program. The KIRIS system reused some items once and others twice. Koretz & Barron explored whether there was a relationship between the mean score gains of schools and the discrepancy in performance between old and reused items. There was no relationship in reading, but in three of four comparisons, there was a modest relationship in mathematics. On average, new items were modestly more difficult, relative to old ones, in higher-gain schools. Their findings most likely underestimate the potential of this approach because they lacked audit items designed for this purpose. Because of similarities between new and old operational items, the effects of inappropriate test preparation may have generalized to some degree to the new items, attenuating the performance difference.

There are a variety of ways in which higher-order differences might be used. As in the Koretz & Barron (1998) study, one might compare schools with particularly rapid gains to others. One might also categorize schools on the basis of knowledge of their test preparation approaches or variables that might be related to test preparation, such as school demographics or initial score level. Particularly in early applications, it may be productive to work in an exploratory direction, first measuring the performance discrepancy and then examining the characteristics of schools at each end of the distribution.

Discussion

The principle underlying the design of SMAs is straightforward. Predictable recurrences in an operational assessment provide opportunities for score inflation because they permit forms of test preparation that undermine the test's representation of the domain. Audit items are designed to avoid these predictable recurrences while still supporting the same intended inferences about student achievement. Discrepancies in performance between operational and audit items provide a measure of possible score inflation. The presence of an audit component may also alter the incentives facing educators and reduce the frequency of inappropriate test preparation and score inflation.

In practice, however, the development of SMAs is likely to be complex. First, it will often be necessary to develop additional specificity about the knowledge and skills that operational items are intended to tap and which portions of the domain they are intended to represent. Second, the developer must identify the gratuitous recurrences that permit inappropriate test preparation and ideally should ascertain which are most used by educators. Finally, the indeterminacy of scale produced by score inflation, particularly in assessments that rely on NEAT linking, creates problems for analysis that are briefly discussed in the preceding section. This indeterminacy will restrict the SMA designs employed, constrain the analyses undertaken, or necessitate collection of additional data. Additional research is needed to evaluate these practical aspects of self-monitoring assessments.

One might ask: why not merely make the entire test sufficiently unpredictable to avoid score inflation? This alternative is likely to be impractical, both because of the

volume of task development it would require and because of technical difficulties it would create, for example, in linking.

Finally, the use of SMAs faces a political barrier. In our current test-based accountability system, every one has the same incentive: to maximize score gains. Over the moderate term, SMAs may encourage the development of greater meaningful, generalizable improvements in student learning, but they can only have a neutral or negative effect on score gains. Thus, employing SMAs will require both changes in policy or the involvement of an independent third party without a stake in maximizing score gains.

References

Ad Hoc Committee on Confirming Test Results. (2001). *Using the National Assessment of Educational Progress to confirm state test results*. Washington, DC: National Assessment Governing Board.

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning, 2*, 67-90.

Dunbar, S. B. (2008). Enhanced assessment for school accountability and student achievement. In 263-273. K. E. Ryan & L. A. Shepard (Eds.), *The Future of Test-Based Educational Accountability.* Mahwah, NJ: Lawrence Erlbaum Associates,

Eduwonkette (2008, July 1). An immodest proposal. Last retrieved from http://blogs.edweek.org/edweek/eduwonkette/2008/07/an_immodest_proposal.html on August 15, 2008.

Feuer, M. J. (2010). Externalities of testing: Lessons from the blizzard of 2010. In Henry I. Braun (Chair), *Innovations in Test-Based Educational Accountability.* Invited symposium presented at the annual meeting of the American Educational Research Association, Denver, CO, May 1.

Figlio, D., & Getzler, L. S. (2006). Accountability, ability and disability: *Advances in Applied Microeconomics, 14*, 35-49.

Fischer, G. H., & Molenaar I.W. (1995). *Rasch models: foundations, recent developments and applications.* New York: Springer-Verlag.

Fitzpatrick, A. R. (2008). The impact of anchor test configuration on student proficiency rates. *Educational Measurement: Issues and Practice*, *27*(4), 34-40.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In: G. H. Fischer, & I. W.Molenaar (eds.). *Rasch models: Foundations, recent developments and applications.* (pp.69-96). New York: Springer-Verlag.

Holcombe, R. W., Jennings, J. L., and Koretz, D. M. (2010). Predictable patterns that facilitate score inflation: a comparison of New York and Massachusetts tests. Unpublished manuscript submitted for review, Harvard University, Cambridge, MA.

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high-stakes testing in the Chicago public schools. *Journal of Public Economics, 89*, 761-796.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*(2), 125–160.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th edition)*, (pp. 18-64). Westport, CT: Praeger.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). *What do test scores in Texas tell us?* (Issue Paper IP-202). Santa Monica, CA: RAND. http://www.rand.org/publications/IP/IP202/.

Koretz, D. (2007). Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian. In Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.), *Linking and Aligning Scores and Scales.* New York: Springer-Verlag, 339-353.

Koretz, D., & Barron, S. I. (1998). *The Validity of Gains on the Kentucky Instructional Results Information System (KIRIS).* MR-1014-EDU, Santa Monica: RAND. http://www.rand.org/publications/MR/MR1014/

Koretz, D., Barron, S., Mitchell, K., & Stecher, B. (1996). *The Perceived Effects of the Kentucky Instructional Results Information System (KIRIS).* MR-792-PCT/FF, Santa Monica: RAND.

Koretz, D., & Hamilton, L. S. (2006). Testing for accountability in K-12. In R. L. Brennan (Ed.), *Educational measurement (4th edition)*, 531-578. Westport, CT: American Council on Education/Praeger.

Koretz, D., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The effects of high-stakes testing: Preliminary evidence about generalization across tests, in R. L. Linn (chair), *The Effects of High Stakes Testing*, symposium presented at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Chicago, April.

Koretz, D., and McCaffrey, D. (2005). *Using IRT DIF Methods to Evaluate the Validity of Score Gains.* CSE Technical Report 660. Los Angeles: Center for the Study of Evaluation, University of California.

Koretz, D., McCaffrey, D., & Hamilton, L. (2001). *Toward a Framework for Validating Gains Under High-Stakes Conditions*. CSE Technical Report 551. Los Angeles: Center for the Study of Evaluation, University of California.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, N.J.: Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement: Third Edition*. New York: American Council on Education/ Macmillan, pp. 13-104.

Morley, M.E., Bridgeman, B., and Lawless, R. R. (2004). *Transfer Between Variants of Quantitative Items*. Princeton , N.J.: Educational Testing Service (GRE Board Research Report No. 00-06R).

Rothstein, R. (2008). *Holding accountability to account: How scholarship and experience in other fields inform exploration of performance incentives in education.* Nashville, TN: National Center on Performance Incentives, Vanderbilt University.

Rubinstein, J. (2002): *The Princeton Review: Cracking the MCAS Grade 10 Mathematics*. New York, NY: Random House.

Schemo, D. J., & Fessenden, F. (2003, December 3). Gains in Houston schools: How real are they? *New York Times* .

Stecher, B. (2002). Consequences of large-scale, high-stakes testing on school and classroom practice. In L. Hamilton, et al., *Test-based Accountability: A Guide for Practitioners and Policymakers*. Santa Monica: RAND. http://www.rand.org/publications/MR/MR1554/MR1554.ch4.pdf

Stecher, B. M., & Mitchell, K. J. (1995): Portfolio Driven Reform: Vermont Teachers' Understanding of Mathematical Problem Solving (CSE Technical Report 400). Los Angeles, CA: University of California Center for Research on Evaluation, Standards, and Student Testing.

Van der Linden, W. J., & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer-Verlag.

Figures

Figure 1. An example of an IIS design with NEAT linking

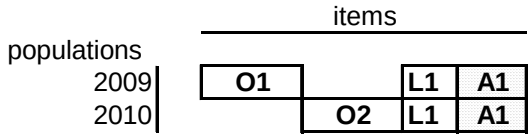


Figure 2. An example of an IID design with NEAT chain linking

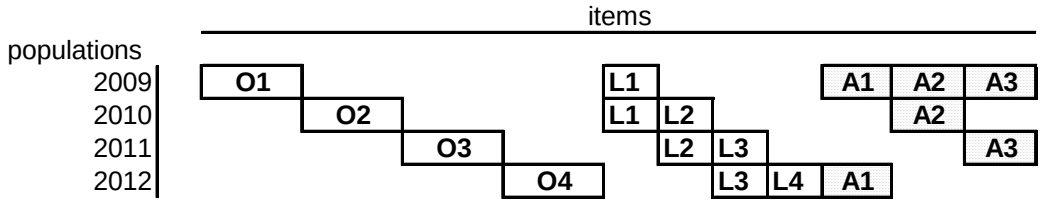


Figure 3. An example of the PIN design with NEAT linking

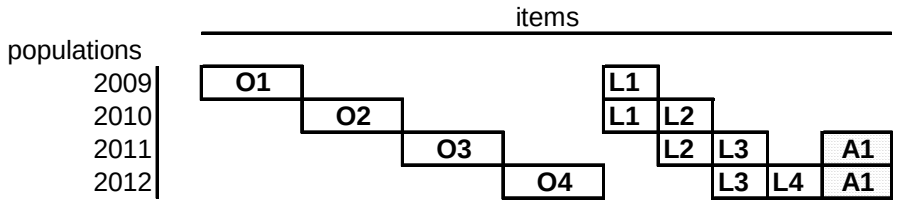


Figure 4. An example of the IPIN design with NEAT linking

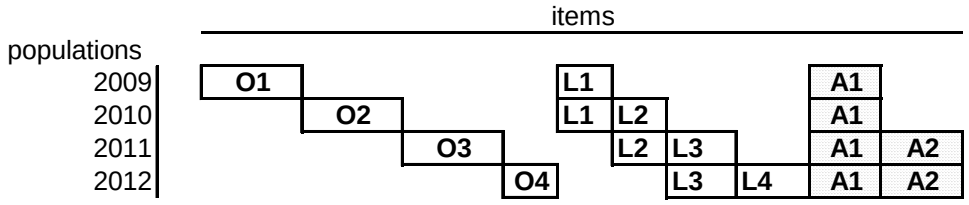
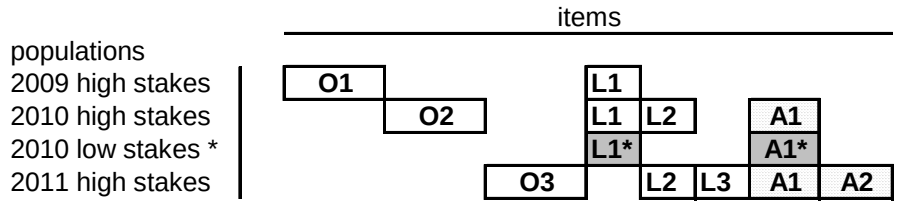


Figure 5. An example of a PIEC design with NEAT linking



* administered to students in a different jurisdiction